1. Text mining
2. Preparation:
3. Remove tweet id and coordinate type because these two attribute have no affects for prediction on place.
4. Impute the missing data of country and place with k-NN method.
5. Divide coordinates into two attributes which are latitude and longitude and set both of their type as numerical.
6. Replace the missing value of latitude and longitude with average value group by place name.
7. Aggregate the average value of latitude and longitude for each place
8. Replace the missing value with the average value
9. Considering there are some places only have one example, their average value is not able to be calculated, therefore, these remaining missing values were removed.
10. Unify the language of “Country” as English, and found there are only three distinct countries: Spain, Portugal, and France.
11. Select prediction method:
12. Naïve Bayes was not chosen based on the Assumption: attributes are independent
13. Decision tree & rule based: relative low accuracy
14. SVM: suit for binary classification.
15. KNN:

Sensitive to noisy features but this disadvantage can be overcome by adjusting the weights for each attribute;

Although it Cause large cost on space and time, it still have better performance compared with other method.

1. Linear regression: only suit for the conditions when all the attributes that used to predict are numerical type.
2. Neural network:

Although it require a powerful HW, it doesn’t need to be re-programmed and powerful to sole classification and other problems.